Modeling Frequency and Severity of Claims with the Generalized Linear Transformed Cluster-Weighted Model

N. Počuča, T. Miljkovic, P. Jevtić, P. McNicholas

July 2, 2018

Abstract

In this paper, we propose a generalized linear cluster-weighted transformed model (GL-TCWM) that allows for modeling non-Gaussian distribution of the continuous covariates. Additionally, our zero-inflated GL-TCWM allows for modeling zero-inflated cluster weighted distribution of claims that is more suitable in the insurance applications. We describe an expectation-optimization (EM) algorithm for parameter estimation in GL-TCWM. Cluster-weighted models are considered as a flexible family of mixture models for fitting the joint distribution of a random vector composed of a response variable and a set of continues and discrete covariates. However, these models have a few limitations when it comes to the insurance applications and may provide suboptimal results. A simulation study showed that the GL-TCWM performs well for different settings in contrast to the existing mixture-based approaches. A real data set based on French auto-mobile policies is used to illustrate the application of the proposed model.

KEY WORDS: finite mixture models, GLM, GL-TCWM, CWM, ratemaking, automobile claims.
JEL CLASSIFICATION: C02, C40, C60.

1 Introduction

Predictive modeling gained a lot of attention in the past decade in the area of actuarial science, risk management, and insurance in general. While the term predictive modeling has been used in many other areas, in context of insurance, it is referred to as a process of leveraging statistics in estimating the insurance cost (see Frees and Meyers, 2014). Various predictive models are used in the area of actuarial science with generalized linear models (GLMs) being the most popular tools actively integrated in pricing, reserving, and underwriting

of property and casualty insurance. The most recent extensions of the GLM models proposed by Garrido et al. (2016) and Shi et al. (2015) allow for relaxing the assumption of independence between number and size of claims. Several GLM extensions based on copulas have been considered (e.g., Frees et al., 2016; Krämer et al., 2013; Czado et al., 2012; Frees and Wang, 2006).

Ingrassia and Minotti (2015) proposed a cluster-weighted models (CWMs) as a flexible family of mixture models for fitting the joint distribution of a random vector composed of a response variable and a set of mixed-type covariates with the assumption that continues covariates come from Gaussian distribution. In this paper paper, we consider two extensions of CWM model to allow for a transformation of the distribution of non-Gaussian continues covariates and to allow for modeling a zero-inflated Poisson (ZIP) claims distribution. We define our proposed model as cluster-weighted transformed model or GL-TCWM which is more suitable for the insurance applications. The CWM models with Gaussian assumptions have been proposed by Gershenfeld (1997), Gershenfeld and Metois (1999), and Gershenfeld (1999) in a context of media technology. Some extensions of this class of models have been considered by Punzo and Ingrassia (2014), Ingrassia and Punzo (2014), and Ingrassia and Vittadini (2014).

There is a growing interest in modeling insurance losses using mixture models. Several recent mixture models of univariate insurance data have been developed, including work by Lee and Lin (2010), Verbelen et al. (2015), and Miljkovic and Grün (2016). A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking was studied by Bermúdez and Karlis (2012). The authors used the EM algorithm to determine the number of components in the mixture. Another Poisson mixture model for count data was considered by Brown and Buckley (2015) with application in managing a Group Life insurance portfolio. Motivated by the idea of mixture modeling, we believe that the extension of the univariate mixture modeling can be applied to mixture modeling of regressions including the GLMs, where losses are modeled as a function of several covariates.

This paper is organized as follows. Section 2 presents the proposed model for mixture of GLMs. Section 3 applies the proposed model on a real data of French automobile claims. An extensive simulation study is discussed in Section 4. Conclusion is provided in Section 5.

2 Proposed Model

2.1 Background

Let $(\boldsymbol{X}', Y)'$ be the pair of a vector of covariates $\boldsymbol{X}$ and a response variable Y. Assume this set is defined on some space Ω that takes values in appropriate Euclidian subspace. Further, assume that there exists G partitions

of Ω, denoted as $\Omega_1, \ldots, \Omega_G$. Gershenfeld (1997) characterized the Cluster weighted models as a finite mixture of GLMs hence, the joint distribution $f(\boldsymbol{x}, y)$ of $(\boldsymbol{X}', Y)'$ is expressed as follows

$$f(\boldsymbol{x}, y) = \sum_{j=1}^{G} \tau_j f(y|\boldsymbol{x}; \Omega_j) f(\boldsymbol{x}; \Omega_j). \tag{2.1}$$

The pair $f(y|\boldsymbol{x}; \Omega_j)$ and $f(\boldsymbol{x}; \Omega_j)$ are conditional and marginal distributions of $(\boldsymbol{X}', Y)'$ respectively, while τ_j represents the weight of the jth component such that $\sum_{j=1}^{G} \tau_j = 1$, $\tau_j > 0$. Ingrassia and Minotti (2015) proposed a flexible family of mixture models for fitting the joint distribution of a random vector $(\boldsymbol{X}', Y)'$ by splitting the covariates into continues and discrete as $\boldsymbol{X} = (\boldsymbol{V}', \boldsymbol{W}')'$. This assumption of independence between continues and discrete covariates allows us to multiply their corresponding marginal distributions. Thus, for this setting the model in (2.1) is reformulated as follows

$$f(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{G} \tau_j f(y|\boldsymbol{x}; \boldsymbol{\vartheta_j}) f(\boldsymbol{x}; \boldsymbol{\theta_j}) = \sum_{j=1}^{G} \tau_j f(y|\boldsymbol{x}; \boldsymbol{\vartheta_j}) f(\boldsymbol{v}\boldsymbol{\theta_j^\star}) f(\boldsymbol{w}; \boldsymbol{\theta_j^{\star\star}}) \tag{2.2}$$

where $\boldsymbol{v}$ and $\boldsymbol{w}$ are the vectors of continues and discrete covariates respectively, the $f(y|\boldsymbol{x}; \boldsymbol{\vartheta_j})$ is a conditional density of $Y|\boldsymbol{x}$, with parameter vector $\boldsymbol{\vartheta_j}$, the $f(\boldsymbol{v}; \boldsymbol{\theta_j^\star})$ is the marginal distribution of $\boldsymbol{v}$ with parameter vector $\boldsymbol{\theta_j^\star}$. the $f(\boldsymbol{w}; \boldsymbol{\theta_j^{\star\star}})$ is the marginal distribution of $\boldsymbol{w}$ with parameter vector $\boldsymbol{\theta_j^{\star\star}}$. Finally $\boldsymbol{\Phi} := (\boldsymbol{\theta^\star}, \boldsymbol{\theta^{\star\star}}, \boldsymbol{\tau}, \boldsymbol{\vartheta})$ includes all model parameters. In addition, the conditional distribution $f(y|\boldsymbol{x}; \boldsymbol{\vartheta_j})$ is assumed to belong to an exponential family of distributions and as such can be modeled in the framework of GLMs. Here, the marginal distribution of continues covariates is assumed to be Gaussian type. Unfortunately, this assumption is too strong for using insurance related applications specifically in rate-making. To relax this constraint, we develop a new model that allows for modelling of non-Gaussian covariates as discussed in the next section.

2.2 Generalized linear transformed cluster-weighted model (GL-TCWM)

We proceed to extend (2.2) by splitting the continues covariates further as $\boldsymbol{V} := (\boldsymbol{U}', \boldsymbol{T}')'$, where $\boldsymbol{U}$ is a set of non-Gaussian covariates, and $\boldsymbol{T}$ as a set of Gaussian covariates. Here, from the perspective of non-Gaussian form, the $\boldsymbol{U}$ is considered to be an untransformed set of covariates. The transformation function $\phi(\cdot)$ is applied to any non-Gaussian covariates and will represent a map to Gaussian form. All this, for the purposes of transforming non-Gaussian covariates into Gaussian ones. Thus (2.2) is now recovered as

$$f(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{G} \tau_j f(y|\boldsymbol{x}; \boldsymbol{\vartheta_j}) f(\boldsymbol{t}; \boldsymbol{\theta_j^\star}) f(\boldsymbol{w}; \boldsymbol{\theta_j^{\star\star}}) \phi(f(\boldsymbol{u}; \boldsymbol{\theta_j^{\star\star\star}})) \tag{2.3}$$

where $f(\boldsymbol{t};\boldsymbol{\theta_j^\star})$ represents the marginal density of Gaussian covariates, with parameter vector $\boldsymbol{\theta^\star}$, and $f(\boldsymbol{u};\boldsymbol{\theta_j^{\star\star\star}})$ as the marginal density of the non-Gaussian covariates with parameter vector $\boldsymbol{\theta_j^{\star\star\star}}$.

2.2.1 Transformation Function Selection - Lognormal Distribution

The choice of a transformation function is limited to a set of bijective maps from a non-Gaussian to Gaussian density form. As it is relevant to most actuarial applications in this paper, we focus on the log-normal distribution for non-Gaussian covariates. Hence, with our log-normal assumption for $f(\boldsymbol{u};\boldsymbol{\theta_j^{\star\star\star}})$, we have that $\boldsymbol{u}$ is defined on $\mathbb{R}^+$ with parameter vector $\boldsymbol{\theta_j^{\star\star\star}}$, and having marginal density

$$f\left(\boldsymbol{u};\boldsymbol{\theta_j^{\star\star\star}} := (\boldsymbol{\mu_j},\boldsymbol{\Sigma_j})\right) = \frac{1}{(\prod_{i=1}^{N} u_i)|\boldsymbol{\Sigma_j}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln \boldsymbol{u}-\boldsymbol{\mu_j})^{'}\boldsymbol{\Sigma_j^{-1}}(\ln \boldsymbol{u}-\boldsymbol{\mu_j})\right]. \tag{2.4}$$

In this case, by applying a particular function (see Appendix 7.3), we have that we have that

$$\phi(f(\boldsymbol{u};\boldsymbol{\theta_j^{\star\star\star}})) = f(\boldsymbol{u};\boldsymbol{\theta_j^{\star\star\star}})\prod_{i=1}^{N} u_i =: f_n(\ln \boldsymbol{u},\boldsymbol{\theta_j^{\star\star\star}}), \tag{2.5}$$

where $f_n(\ln \boldsymbol{u},\boldsymbol{\theta_g^{\star\star\star}})$ is the desired Gaussian form. With the covariate $\boldsymbol{u}$ now transformed appropriately we can rewrite (2.3) as

$$f(\boldsymbol{x},y;\Phi) = \sum_{j=1}^{G}\tau_j f(y|\boldsymbol{x};\boldsymbol{\vartheta_j})f(\boldsymbol{t};\boldsymbol{\theta_j^\star})f(\boldsymbol{w};\boldsymbol{\theta_j^{\star\star}})f_n(\ln \boldsymbol{u},\boldsymbol{\theta_j^{\star\star\star}}) \tag{2.6}$$

2.3 Zero inflated GL-TCWM

For the zero-inflated Poisson model (ZIP model see (Lambert, 1992)) we can split the conditional density $f(y|\boldsymbol{x},\boldsymbol{\vartheta_j})$ into zero and non-zero densities. The response y variable when $y=0$ are distributed with density $f(0|\boldsymbol{x},\boldsymbol{\vartheta_j})$. The response y values when $y>0$ are distributed with density $f(y>0|\boldsymbol{x},\boldsymbol{\vartheta_j})$. Given the conditional density now defined for the ZIP model, (2.3) can be re-written as follows

$$f(\boldsymbol{x},y) = \sum_{j=1}^{G}\tau_j\left[f(0|\boldsymbol{x};\boldsymbol{\vartheta_j}) + f(y>0|\boldsymbol{x};\boldsymbol{\vartheta_j})\right]f(\boldsymbol{t};\boldsymbol{\theta_j^\star})f(\boldsymbol{w};\boldsymbol{\theta_j^{\star\star}})\phi(f(\boldsymbol{u};\boldsymbol{\theta_j^{\star\star\star}})). \tag{2.7}$$

Let $\tilde{\boldsymbol{X}} = [1,\boldsymbol{x}]$, where $\tilde{\boldsymbol{X}}$ is a vector of covariates with the addition of a placeholder for the intercept in the GLM. We denote the Poisson conditional density as $f^P(y|\boldsymbol{x};\boldsymbol{\beta_j})$, where $y\in\{0,1,\dots\}$, and $\boldsymbol{\beta_j}$ is the row

coefficient vector. Here, the link function will be modelled with log-link for the GLM:

$$\lambda_j = e^{\tilde{\boldsymbol{X}}\boldsymbol{\beta}_j'}, \qquad\qquad f^P(y|x;\lambda_j) = e^{-\lambda_j}\frac{{\lambda_j}^y}{y!}.$$

Next, we introduce a Bernoulli process for the conditional density. We denote the density as $f^B(y|\boldsymbol{x};\bar{\boldsymbol{\beta}}_{\boldsymbol{j}})$, where $y \in \{0,1\}$, and $\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}$ as the coefficient vector. Here, the GLM will be modeled with the associated logit link function

$$\psi_j = \frac{e^{\tilde{\boldsymbol{X}}\bar{\boldsymbol{\beta}}_j{}'}}{1+e^{\tilde{\boldsymbol{X}}\bar{\boldsymbol{\beta}}_j{}'}}, \qquad\qquad f^B(y|\boldsymbol{x};\bar{\boldsymbol{\beta}}_{\boldsymbol{j}}) = \begin{cases} \psi_j, & y=0 \\ 1-\psi_j, & y=1 \end{cases}$$

Now, given a combination of two preceding models, we introduce the ZIP process in which zero counts come from two random variables. One comes from Bernoulli random variable which generates structural zeros, and the other comes from the Poisson random variable. The coefficients $\boldsymbol{\vartheta}_{\boldsymbol{j}} = \{\boldsymbol{\beta}_{\boldsymbol{j}}, \bar{\boldsymbol{\beta}}_{\boldsymbol{j}}\}$ correspond to the two above introduced conditional densities where the coefficients are estimated using a generalized linear model as in Lambert (1992). The components of ZIP conditional density $f(y|\boldsymbol{x};\boldsymbol{\vartheta}_{\boldsymbol{j}})$ are

$$f(0|x;\boldsymbol{\vartheta}_{\boldsymbol{j}}) = \psi_j + (1-\psi_j)e^{-\lambda_j} \qquad \text{and} \qquad f(y>0|x;\boldsymbol{\vartheta}_{\boldsymbol{j}}) = (1-\psi_j)e^{-\lambda_j}\frac{\left(e^{-\lambda_j}\right)^y}{y!}.$$

Also, the link functions to consider are log-link for the Poisson process and logit link for the Bernoulli

$$\psi_j = \frac{e^{\tilde{\boldsymbol{X}}\bar{\boldsymbol{\beta}}_j{}'}}{1+e^{\tilde{\boldsymbol{X}}\bar{\boldsymbol{\beta}}_j{}'}} \qquad \text{and} \qquad \lambda_j = e^{\tilde{\boldsymbol{X}}\boldsymbol{\beta}_{\boldsymbol{j}}'}.$$

Let paramater ψ_j denote the probability that the zero comes from the Bernoulli distribution of jth component, and the parameter λ_j characterizes the jth Poisson distribution. This allows for a more nuanced approach to handling the inflation of zeros similarly as in Bermúdez and Karlis (2012).

3 Introducing Bernoulli-Poisson Partitioning

The single component ZIP model assumes that the inflated zeros emanate from both a Bernoulli and Poisson random variables. The non-zeros are assumed to come exclusively from the Poisson random variable. However, recent research extends the single component ZIP models to mixture models for heterogeneous count data with excess zeros (see (Bermúdez and Karlis, 2012)). In mixtures of ZIPs, zeros are assumed to come from multiple different Binomial and Poisson random variables. Difficulties are apparaent during the

maximization step of the EM when means of covariates are very close together (see Lim et al. (2014)) . However, misclassification error can be reduced using parsimonious models such as McNicholas et al. (2010). In this work, we propose a new method rectify this problem and partition the dataset using Bernoulli and Poisson CWMs. The ZIP model is a combination of both these processes, CWM can accurately estimate the initialization of the EM algorithm for ZIP. The work of Lambert (1992) specifies that the MLE estimates for coefficients provide an excellent guess allowing EM to converge quickly. Recall $(\boldsymbol{X}', Y)'$ to be a vector defined on some sample space Ω. As discussed, this sample space is partitioned into G non-overlapping sets such that their union constitutes this sample space ie. $\Omega = \bigcup_{i=1}^{G} \Omega_i$. However, contingent on a model choice each particular set Ω_i may take a different shape. Specifically, if we introduce the Bernoulli model in a generalized form for conditional density (see Ingrassia and Minotti (2015) for specific cases), we have the sample space Ω^B and joint pdf f^B to be

$$\Omega^B = \bigcup_{i=1}^{G} \Omega_i^B \qquad \text{and} \qquad f^B(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{G} \tau_j f^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_{\boldsymbol{j}}) f(\boldsymbol{t}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star}) f(\boldsymbol{w}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star}) \phi(f(\boldsymbol{u}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star\star})).$$

Similarly if we introduce a Poisson model in a generalized form the sample space Ω^P and joint pdf f^P become

$$\Omega^P = \bigcup_{i=1}^{G} \Omega_i^P \qquad \text{and} \qquad f^P(\boldsymbol{x}, y; \Phi) = \sum_{j=1}^{G} \tau_j f^P(y|\boldsymbol{x}; \boldsymbol{\beta}_{\boldsymbol{j}}) f(\boldsymbol{t}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star}) f(\boldsymbol{w}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star}) \phi(f(\boldsymbol{u}; \boldsymbol{\theta}_{\boldsymbol{j}}^{\star\star\star})).$$

Now, construct a new partitioning of a sample space Ω such that

$$\Omega = \Omega^Z = \bigcup_{l,j \in \{1 \ldots G\}} \Omega_{l,j}^Z := \bigcup_{l,j \in \{1 \ldots G\}} \Omega_l^B \cap \Omega_j^P$$

and is now result of a model in which each component is captured by pdf that is of mixture of particular Bernoulli and particular Poisson.

$$f_{l,j}^Z(y|\boldsymbol{x}; \boldsymbol{\beta_{zl}}, \boldsymbol{\beta_j}) = f^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_{\boldsymbol{j}}) + (1 - f^B(y|\boldsymbol{x}; \bar{\boldsymbol{\beta}}_{\boldsymbol{j}})) f^P(y|\boldsymbol{x}; \boldsymbol{\beta_j})$$

The expectation-maximization (EM) algorithm (see Dempster et al. (1977)) is then used to estimate this new mixture of up to G^2 specific ZIP models. The initialization parameters are provided by Bernoulli and Poisson CWMs resulting in parameter pairs (ψ_l, λ_j) where $l, j \in \{1 \ldots G\}$. After estimation procedure the newly constructed ZIP model is then compared against the standard poisson model using a Stats test which is commented in section 3.6.

3.1 The EM Algorithm for Parameter Estimation

In most finite mixture problems, the standard method for estimating the optimal number of components G is based on the EM algorithm and further discussed by McLachlan and Peel (2000). The EM algorithm is based on the maximum likelihood estimation. The initial values of the parameter estimates are generated through a stochastic initialization, then the algorithm proceeds by alternation of the E-step and M-step to update the parameter estimates as well as missing data which in this case correspond to the posterior probability that x_i comes from the jth mixture component, computed at each iteration of the EM algorithm. To find an optimal number of components, maximum likelihood estimation is obtained over a range of G, and the best model is selected based on a chosen model selection criterion.

The convergence of the EM algorithm is achieved when the relative increase in the log-likelihood function is no bigger than a small pre-specified tolerance value or the number of iterations reach a limit. In this subsection, we explain the parameter estimation in line with the CWM methodology proposed by Ingrassia and Minotti (2015). The proposed GL-TCWM model is based on the assumption that $f(y|\boldsymbol{x}, \vartheta_j)$ belongs to the exponential family of distributions that are strictly related to GLMs. The link function relates the expected value $g(\mu_j) = \beta_{0j} + \beta_{1j}x_1 + \cdots + \beta_{pj}x_p$. We are interested in estimation of the vector $\boldsymbol{\beta}_j$, thus the distribution of $y|\boldsymbol{x}, \mathcal{R}_j$ is denoted by $f(y|\boldsymbol{x}, \beta_j, \lambda_j)$, where λ_j denotes an additional parameter to account for when a distribution belong to a two-parameter exponential family.

The marginal distribution $f(\boldsymbol{x}; \theta_j)$ has the following components: $f^{\mathrm{T}}(\boldsymbol{u}; \theta_j^{'})$, $f(\boldsymbol{t}; \theta_j^{''})$, and $f(\boldsymbol{w}; \theta_j^{\star\star\star})$. The first two components are modeled as Gaussian density with mean $\boldsymbol{\mu_j}$ and covariance matrix $\boldsymbol{\Sigma}_j$ as $\phi(\boldsymbol{v}; \boldsymbol{\mu_j}, \boldsymbol{\Sigma}_j)$. The marginal density $f(\boldsymbol{w}; \theta_j^{\star\star\star})$ assume that each finite discrete covariate W is represented as a vector $\boldsymbol{w}^r = (w^{r1}, \ldots, \boldsymbol{w}^{rc_r})^{'}$ where $w^{rs} = 1$ if w_r has the value s, such that $s \in \{1, \ldots, c_r\}$, and $w^{rs} = 0$ otherwise.

$$f(\boldsymbol{w}, \boldsymbol{\gamma_j}) = \prod_{r=1}^{q} \prod_{s=1}^{c_r} (\gamma_{jrs})^{w^{rs}} \tag{3.1}$$

for $j = 1, \ldots, G$, where $\boldsymbol{\gamma_j} = (\gamma_{j1}^{'}, \ldots, \gamma_{jq}^{'})^{'}$, $\boldsymbol{\gamma}_{jr} = (\gamma_{jr1}^{'}, \ldots, \gamma_{jrc_q}^{'})^{'}$, $\gamma_{jrs} > 0$, and $\sum_{s=1}^{c_r} \gamma_{jrs}$, $r = 1, \ldots, q$. The density $f(\boldsymbol{w}, \boldsymbol{\gamma_j})$ represents the product of q conditionally independent multinomial distributions with parameters $\boldsymbol{\gamma}_{jr}$, $r = 1, \ldots, q$.

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be a sample of n independent observations drawn from model in (2.3). For this

sample, the complete data likelihood function, $L(\mathbf{\Phi})$, is given by

$$L_c(\mathbf{\Phi}) = \prod_{i=1}^{n}\prod_{j=1}^{G}\left[\tau_j f(y_i|x_i,\boldsymbol{\beta}_j,\lambda_j)\phi\{f^T(u_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)\}f(t_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)f(w_i,\gamma_j)\right]^{z_{ij}}, \tag{3.2}$$

where z_{ig} is the latent indicator variable with value of $z_{ig}=1$ indicating that observation $(\boldsymbol{x_i}, y_i)$, originated from the jth mixture component and $z_{ij}=0$ otherwise.

By taking the logarithm of (3.2), the complete-data log-likelihood function $\ell_c(\mathbf{\Phi})$ is written by

$$\ell_c(\mathbf{\Phi}) = \sum_{i=1}^{n}\sum_{j=1}^{G} z_{ij}\big[\log(\tau_j) + \log f(y_i|x_i,\boldsymbol{\beta}_j,\lambda_j) + \log\phi\{f^T(u_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)\} + \log f(t_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j) + \log f(w_i,\gamma_j)\big].$$

3.2 E-Step - Partitioning

The E-step does not depend on the form of density, and the latent data only relate to $\boldsymbol{z}$. The posterior probability that $(\boldsymbol{x_i}, y_i)$ comes from the jth mixture component is calculated at the sth iteration of the EM algorithm as

$$\begin{aligned}\tau_{ij}{}^{(s)} &= E[z_{ij}|(\boldsymbol{x_i}, y_i), \mathbf{\Phi}^{(s)}] \\ &= \frac{\tau_j{}^{(s)} f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta_j}^{(s)},\lambda_j{}^{(s)})\phi\{f^{\mathrm{T}}(\boldsymbol{u}_i,\boldsymbol{\mu}_j^{(s)},\mathbf{\Sigma_j}^{(s)})\}f(\boldsymbol{t}_i,\boldsymbol{\mu}_j^{(s)},\mathbf{\Sigma_j}^{(s)})f(\boldsymbol{w}_i,\gamma_j{}^{(s)})}{f(\boldsymbol{x_i}, y_y;\Phi^{(s)})}.\end{aligned}$$

3.3 M-Step - Partitioning

It follows that at the sth iteration, the conditional expectation of (3.1) on the observed data and the estimates from the $(s-1)$th iteration results in

$$\begin{aligned}Q(\mathbf{\Phi}|\mathbf{\Phi}^{(s-1)}) &= \sum_{i=1}^{n}\sum_{j=1}^{G}\tau_{ij}^{(s-1)}\big[\log(\tau_j) + \log f(y_i|x_i,\boldsymbol{\beta}_j,\lambda_j) + \log\phi\{f^{\mathrm{T}}(u_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)\} \\ &\qquad + \log f(t_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j) + \log f(w_i,\gamma_j)\big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{G}\tau_{ij}^{(s-1)}\log(\tau_j) + \sum_{i=1}^{n}\sum_{j=1}^{G}\tau_{ij}^{(s-1)}\log f(y_i|x_i,\boldsymbol{\beta}_j,\lambda_j) + \sum_{i=1}^{n}\sum_{j=1}^{G}\tau_{ij}^{(s-1)}\log f(t_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j) \\ &\qquad + \sum_{i=1}^{n}\sum_{j=1}^{G}\tau_{ij}^{(s-1)}\log\phi\{f^{\mathrm{T}}(u_i,\boldsymbol{\mu}_j,\mathbf{\Sigma}_j)\} + \sum_{i=1}^{n}\sum_{j=1}^{G}\tau_{ij}^{(s-1)}\log f(w_i,\gamma_j).\end{aligned}$$

The M-step requires maximization of the Q-function with respect to $\mathbf{\Phi}$ which can be done separately for each

term on the right hand side in Equation 2.11. As a result, the parameter updates $\hat{\tau}_j$, $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\sigma}}_j$, and $\hat{\boldsymbol{\gamma}}_j$ on the $(s+1)$th iteration are:

$$\hat{\tau}_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ij}^{(s)}, \qquad \hat{\boldsymbol{\mu}}_j^{(s+1)} = \frac{1}{\sum_{i=1}^{n}\tau_{ij}^{(s)}}\sum_{i=1}^{n}\tau_{ij}^{(s)}\boldsymbol{t}_i, \qquad \hat{\boldsymbol{\gamma}}_{jr}^{(s+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}^{(s)}\omega_i^{rs}}{\sum_{i=1}^{n}\tau_{ij}^{(s)}},$$

$$\hat{\boldsymbol{\sigma}}_j^{(s+1)} = \frac{1}{\sum_{i=1}^{n}\tau_{ij}^{(s)}}\sum_{i=1}^{n}\tau_{ij}^{(s)}(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})(\boldsymbol{t}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})',$$

The log-normal distribution is relevant for modelling actuarial data. Parameter estimates for the log-normal distribution follow similar suit.

$$\hat{\tau}_j^{(s+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ij}^{(s)}, \qquad \hat{\boldsymbol{\mu}}_j^{(s+1)} = \frac{1}{\sum_{i=1}^{n}\tau_{ij}^{(s)}}\sum_{i=1}^{n}\tau_{ij}^{(s)}\ln\boldsymbol{u}_i, \qquad \hat{\boldsymbol{\gamma}}_{jr}^{(s+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}^{(s)}\omega_i^{rs}}{\sum_{i=1}^{n}\tau_{ij}^{(s)}},$$

$$\hat{\boldsymbol{\sigma}}_j^{(s+1)} = \frac{1}{\sum_{i=1}^{n}\tau_{ij}^{(s)}}\sum_{i=1}^{n}\tau_{ij}^{(s)}(\ln\boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})(\ln\boldsymbol{u}_i - \hat{\boldsymbol{\mu}}_j^{(s+1)})'.$$

The estimates of $\boldsymbol{\beta}$ are computed by maximizing each of the G terms

$$\sum_{i=1}^{n}\tau_{ij}^{(s)}\log f(y_i|\boldsymbol{x}_i,\boldsymbol{\beta}_j,\lambda_j). \tag{3.3}$$

Maximization of (3.3) is performed by numerical optimization in R software in a similar framework the mixture of generalized linear models are implemented. For additional details about this implementation the reader is refer to Wedel and De Sabro (1995) and Wedel (2002). For insurance applications, current TCWM model can be used for modeling frequency of claims assuming that $\boldsymbol{Y}$ belongs to Poisson or Bernoulli distributions. When modelling severity of claims, $\boldsymbol{Y}$ can be assumed accommodate Gamma or Lognormal distributions. All of these applications are based on CWM as the underlying approach. For additional information, the reader is referred to the manual of the `flexCWM` package manual for **R** users written by Mazza A., Punzo A., and Ingrassia S. (2015).

3.4 E-step - ZIP Model

The optimization of the zero-inflated model uses the EM algorithm to maximize the complete-data log-liklihood (Lambert, 1992). Using current estimates $\boldsymbol{\vartheta_k} = \{\psi_k, \lambda_k\}$ from the partition Ω_{Z_k}, we calculate the expected value of z_{ik} by its posterior mean $z_{ik}^{(s)}$ for each cluster, at iteration s. We note that for s = 1 we

use the parameters $\vartheta_k = \{\psi_l, \lambda_j\}$, as the initialization parameters. The E-Step is now calculated as:

$$z_{ik}^{(s)} = \left[1 + \frac{e^{-\psi_k}}{(1+e^{\lambda_k})^{-n}}\right]^{-1} \qquad y_{ik} = 0 \qquad\qquad z_{ik}^{(s)} = 0 \qquad y_{ik} > 0$$

3.5 M-Step - ZIP Model

The M-Step can be split into the maximization of two complete data log-likelihoods using parameters $\boldsymbol{\vartheta}^{(s)}$ calculated from the previous iteration (s). This type of form has been recently solved in Lambert (1992),We decompose $C_T(\boldsymbol{\vartheta}_{\boldsymbol{k}}^{(s+1)}$ into Lambert's equations:

$$C_T(\boldsymbol{\vartheta}_{\boldsymbol{k}}^{(s+1)}; \boldsymbol{\vartheta}_{\boldsymbol{k}}^{(s)}) = \sum_{i=1}^{n}\left(z_{ik}^{(s)}\log(\psi_k) - \log(1+\psi_k) + (1-z_{ik}^{(s)})(y_i \log\lambda_k) - (1-z_{ik}^{(s)})\log y_i!\right)$$

$$C_z(\boldsymbol{\vartheta}_{\boldsymbol{k}}^{(s+1)}; \boldsymbol{\vartheta}_{\boldsymbol{k}}^{(s)}) = \sum_{i=1}^{n}\left(z_{ik}^{(s)}\log(\psi_k) - \log(1+\psi_k)\right) \tag{3.4}$$

$$C_n(\boldsymbol{\vartheta}_{\boldsymbol{k}}^{(s+1)}; \boldsymbol{\vartheta}_{\boldsymbol{k}}^{(s)}) = \sum_{i=1}^{n}\left((1-z_{ik}^{(s)})(y_i \log\lambda_k) - (1-z_{ik}^{(s)})\log y_i!\right) \tag{3.5}$$

The maximization of (3.4) for GLM coefficients λ_k can be found by using a weighted, log-linear Poisson regression with weights $1 - z_{ik}^{(s)}$ [McCullagh and Nelder 1989].While the GLM coefficients of ψ_k can be maximized over a gradient (Lambert, 1992).

3.6 Comparing Models

Once the ZIP models have been found, we wish to compare a single component CWM Poisson model on the same Ω_{Z_k}. We will use Vuong's test for comparing ZIP models to other non-nested count data models. This is fairly standard shown in Ezzahid (2012). Similarly we define $p_n(y_i|x_i)$ to be the probabilities of observed counts using model n. We then denote r_i as the log-ratio of probabilities between the CWM/ZIP, and Vuong's test for hypothesis with $\mathbb{E}[r_{ik}] = 0$ defined:

$$r_{ik} = \log\left(\frac{P_{Z_k}(y_{ik}|x_{ik})}{P_{CWM_k}(y_{ik}|x_{ik})}\right) \qquad\qquad V_k = \frac{\sqrt{n_k}\left(\frac{1}{n_k}\sum_{i=1}^{n_k} r_{ik}\right)}{\sqrt{\frac{1}{n_k}\sum_{i=1}^{n_k}(r_{ik} - \overline{r_k})^2}}$$

Now,V_k is asymptotically normally distributed we have with a 5% significance level the following cut offs:

$V < -1.96$	$\|V\| < 1.96$	$1.96 < V$
CWM-P	Equal*	ZIP

* We note that when models are equal, our heuristic is to choose the simple CWM-P model.

4 Application

4.1 Data

We illustrate the proposed methodology on French motor claims data set by policy. This data set is available as part of the R package `CASdatasets` and it is previously used by Jean-Philippe Boucher and Arthur Charpentier and referenced in Charpentier (2014). The authors demonstrated various GLM modeling approaches for fitting frequency and severity of this data. The claim count including zero claims for 413,169 motor third-party liability policies are provided with the associated risk characteristics. The loss amounts by policy ID are also provided. The amount of loss is modeled as a function of the following covariates: density, car age, driver age, exposure, gas type, car brand, and region.

Table 1: Description of variables of interest

Attribute	Description
Policy ID	Unique identifier of the policy holder
Claim Nb	Number of claims during exposure period (0,1,2,3,4)
Exposure	The exposure of policy in years (0–1.5)
Power	Power level of car ordered categorical (12 levels)
Car Age	Car age in years (5 levels)
Driver Age	Age of a legal driver
Brand	Car brands (7 types)
Gas	Diesel or Regular
Region	Regions in France (10 classifications)
Density	Number of inhabitants per km^2
Loss Amount	Portion of claim the insurance policy pays

4.2 Analysis and Results

4.2.1 Modelling Severity

In this section, we show the results from modeling French motor losses. We consider the following covariates: population density, driver age, car age, car power, and regions. The model that was fitted is defined with the

following equation where $\epsilon \sim \mathcal{N}(0, \sigma)$, and

$$\text{AggClaims} \ \sim \ \text{Density + Driver Age + Region + CarAge + Power} \ + \epsilon.$$

Car age is model as categorical variable with five categories: < 5, $5 - 10$, $11 - 20$, $21 - 30$, > 30. Driver age is modelled as a continuous predictor. The shape of the distribution for driver age indicates that Gaussian assumption is reasonable for this left-truncated data. We recognize that, in the actuarial pricing, driver age may be treated as categorical variable with several categories. However, this flexibility is left up to the user to modify the setting. Beginning with the continuous covariate Density, we want to inspect the shape of its univariate data to see if it follows Gaussian distribution. Figure 2

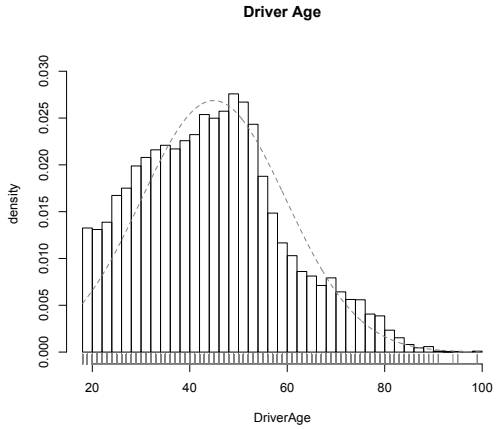

Figure 1: Histogram of driver age with Gaussian distribution fitted to the data.

Table 2: Model comparisons between GL-TCWM and CWM.

Model	G	LL	AIC	BIC
GL-TCWM	1	-118901	237869	238129
	2	-108293	216719	217231
	3	**-107972**	**216143**	**216907**
	4	-107822	215931	216947
CWM	1	-239601	479298	79529
	2	-226334	452921	453433
	3	-216255	432710	433474
	4	**-212452**	**425170**	**426186**

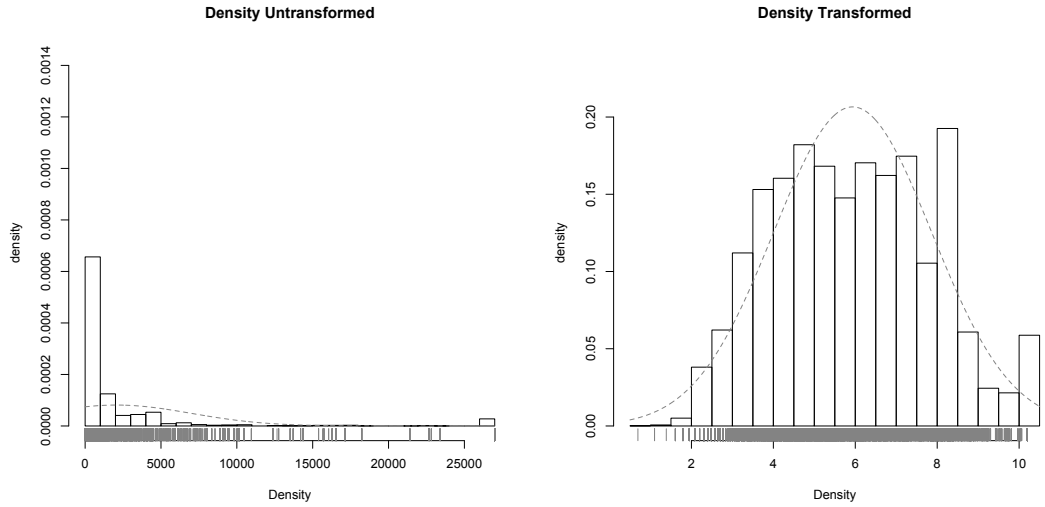

Figure 2: Density variable: Left figure shows the fit when Gaussian distribution is imposed (CMW approach) to highly skewed data. Right figure shows the fit when log-normal transformation is applied (GL-TCWM approach).

4.2.2 Driver Analysis of Severity Model

We now investigate the results of GL-TCWM in relation to the valuation of risk. For practical uses, finding clusters allows us to create different classifications of risk for various fields of drivers. The following model uses GL-TCWM to cluster different drivers in groups allowing one to assign different rates to different clusters. Again, we use the French motor claims data to create a model with claims as the dependent variable and the canonical log-link as the link function.

Table 3: Size of clusters for the GL-TCWM model.

1	2	3	4
7119	1783	2428	4060
Red	Green	Teal	Blue

After fitting the model, we then take a look at the size of each cluster. The GL-TCWM function has chosen four components as the best model to represent the data. The size of each cluster is displayed in Table 3. Attention is brought to largest quantity of drivers that are grouped into cluster 1. This accounts for 46% of all drivers and is fairly concentrated in both plots in Figure 3.3. From these results we can create an insurance model with the following characteristics. Cluster 1 drivers have low variability in claims, thus can be insured with a lower upper limit than the other drivers. From a risk management perspective this is the most ideal case

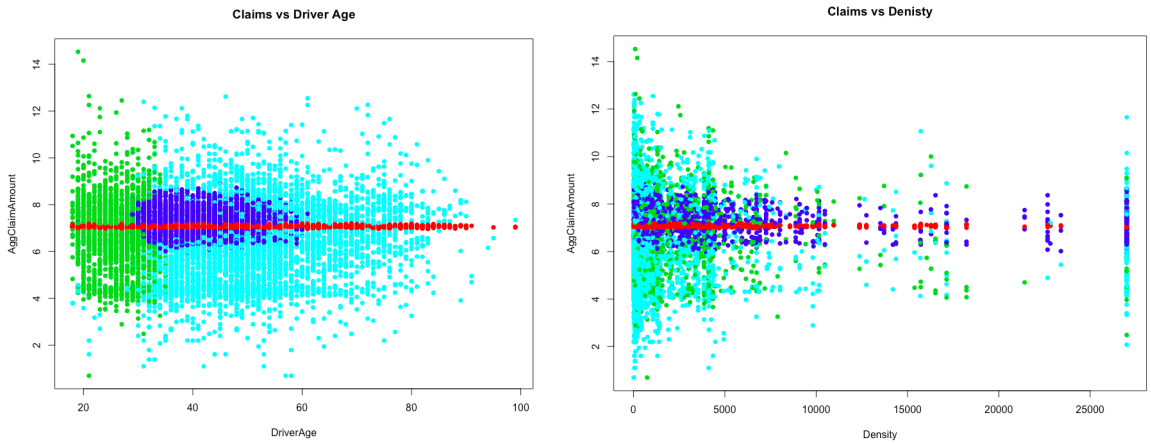

Figure 3: Claims vs Driver Age: The left figure shows the clusters of claims with respect to driver age as the independent variable . The second figure shows the clusters of claims with respect to Density as the independent variable).

as these drivers will claims with very low variance. Cluster 2 drivers have the second lowest variability, thus we would increase the upper limit. The same can be done for the final two Clusters while adjusting the limits accordingly.

Table 4: Summarized volatility information of each cluster.

Volatility Level - (Cluster)	Characteristic	Minimum	Mean	Maximum	$\sigma(0.05)$
V1 - (1)	Aggregate Claims	1035	1172	1340	**52.64**
	Driver Age	18	46.21	99	15.3
	Density	3	1839	27000	4546.906
V2 - (4)	Aggregate Claims	21	1823	7511	**1047.06**
	Driver Age	19	42.06	62	7.2
	Density	3	3407	27000	6618.938
V3 - (3)	Aggregate Claims	2	3109	301300	**13468.54**
	Driver Age	18	52.72	99	12.6
	Density	2	1593	27000	4310.153
V4 - (2)	Aggregate Claims	2	5043	2037000	**59750.99**
	Driver Age	18	25.79	35	4
	Density	6	2219	27000	4590.053

Table 4 shows a breakdown of the types of drivers, ordered by volatility in descending order. Beginning with cluster 1, drivers have a mean age of 46 years. The age of these drivers are fairly spread with a standard deviation of 15 years. However, the one noticeable difference is that these drivers tend to have claims between 1035 to 1340, with a standard deviation of 52.64, and a mean of 1172. That means that these drivers rarely exceed costs and tend to have very low volatility. Which implies that claims tend to be the same across all

Table 5: The coefficients and significance of each cluster.

	V1	(Red)		V2	(Blue)		V3	(Teal)		V4	(Green)	
Coef	Estimate	Error	P	Estimate	Error	P	Estimate	Error	P	Estimate	Error	P
(Intercept)	$7.07 \cdot 10^{0}$	1.00	***	$6.96 \cdot 10^{0}$	1.05	***	$5.92 \cdot 10^{0}$	1.12	***	$8.91 \cdot 10^{0}$	1.13	***
Density	$1.14 \cdot 10^{-4}$	1.00		$-7.03 \cdot 10^{-3}$	1.00	*	$-3.47 \cdot 10^{-2}$	1.01	***	$3.03 \cdot 10^{-2}$	1.01	***
DriverAge	$1.03 \cdot 10^{-4}$	1.00	***	$9.89 \cdot 10^{-4}$	1.00	.	$1.88 \cdot 10^{-2}$	1.00	***	$-6.39 \cdot 10^{-2}$	1.00	***
R23	$4.94 \cdot 10^{-3}$	1.00		$-1.39 \cdot 10^{-1}$	1.04	***	$1.47 \cdot 10^{-1}$	1.12		$-1.91 \cdot 10^{-1}$	1.11	.
R24	$-1.43 \cdot 10^{-2}$	1.00	***	$6.58 \cdot 10^{-2}$	1.02	***	$-3.61 \cdot 10^{-1}$	1.05	***	$-1.71 \cdot 10^{-2}$	1.04	
R25	$-1.43 \cdot 10^{-2}$	1.00	***	$-1.78 \cdot 10^{-1}$	1.04	***	$-1.05 \cdot 10^{-1}$	1.09		$1.59 \cdot 10^{-1}$	1.09	.
R31	$-2.32 \cdot 10^{-3}$	1.00		$2.43 \cdot 10^{-3}$	1.03		$1.98 \cdot 10^{-1}$	1.07	**	$-8.98 \cdot 10^{-2}$	1.06	.
R52	$-1.52 \cdot 10^{-2}$	1.00	***	$2.23 \cdot 10^{-1}$	1.02		$-3.89 \cdot 10^{-1}$	1.06	***	$1.36 \cdot 10^{-1}$	1.06	*
R53	$-1.38 \cdot 10^{-2}$	1.00	***	$2.22 \cdot 10^{-1}$	1.02	***	$-1.78 \cdot 10^{-1}$	1.06	**	$8.09 \cdot 10^{-2}$	1.06	
R54	$-1.47 \cdot 10^{-2}$	1.00	***	$8.69 \cdot 10^{-4}$	1.03	***	$-4.04 \cdot 10^{-1}$	1.08	***	$1.85 \cdot 10^{-1}$	1.07	**
R72	$-5.98 \cdot 10^{-3}$	1.00	***	$1.44 \cdot 10^{-1}$	1.03	***	$-6.77 \cdot 10^{-2}$	1.07		$2.37 \cdot 10^{-2}$	1.06	
R74	$-1.82 \cdot 10^{-2}$	1.00	***	$1.44 \cdot 10^{-1}$	1.06		$-1.05 \cdot 10^{-1}$	1.14		$-3.32 \cdot 10^{-1}$	1.14	*
PE	$-1.21 \cdot 10^{-3}$	1.00		$3.31 \cdot 10^{-2}$	1.02	***	$6.47 \cdot 10^{-2}$	1.05		$-1.83 \cdot 10^{-1}$	1.04	***
PF	$-2.56 \cdot 10^{-3}$	1.00	*	$7.85 \cdot 10^{-2}$	1.02	.	$1.43 \cdot 10^{-1}$	1.04	**	$-3.93 \cdot 10^{-2}$	1.04	
PG	$-5.31 \cdot 10^{-4}$	1.00		$1.83 \cdot 10^{-1}$	1.02	***	$1.35 \cdot 10^{-1}$	1.05	**	$-7.17 \cdot 10^{-2}$	1.04	.
PH	$3.89 \cdot 10^{-3}$	1.00	*	$9.02 \cdot 10^{-2}$	1.03	***	$1.15 \cdot 10^{-1}$	1.07	.	$-1.83 \cdot 10^{-1}$	1.07	**
PI	$-2.00 \cdot 10^{-3}$	1.00		$1.20 \cdot 10^{-1}$	1.03	**	$2.04 \cdot 10^{-1}$	1.08	**	$-8.05 \cdot 10^{-2}$	1.07	
PJ	$8.36 \cdot 10^{-3}$	1.00	***	$-5.91 \cdot 10^{-2}$	1.03	***	$2.54 \cdot 10^{-1}$	1.08	***	$-1.61 \cdot 10^{-1}$	1.08	*
PK	$5.63 \cdot 10^{-3}$	1.00	*	$6.86 \cdot 10^{-2}$	1.04		$1.47 \cdot 10^{-1}$	1.09	.	$-1.80 \cdot 10^{-1}$	1.12	
PL	$9.04 \cdot 10^{-3}$	1.00	**	$4.18 \cdot 10^{-1}$	1.06		$2.31 \cdot 10^{-1}$	1.14	.	$3.09 \cdot 10^{-1}$	1.18	.
PM	$-7.52 \cdot 10^{-3}$	1.01		$1.14 \cdot 10^{-1}$	1.08	***	$-7.67 \cdot 10^{-2}$	1.20		$1.68 \cdot 10^{-1}$	1.28	
PN	$-3.92 \cdot 10^{-3}$	1.01		$-5.77 \cdot 10^{-1}$	1.09		$-4.75 \cdot 10^{-1}$	1.24	*	$-1.58 \cdot 10^{-1}$	1.21	
PO	$3.48 \cdot 10^{-3}$	1.01		$3.57 \cdot 10^{-2}$	1.09	***	$1.01 \cdot 10^{-1}$	1.23		$-6.48 \cdot 10^{0}$	1.63	***
C1	$5.67 \cdot 10^{-3}$	1.00	*	$1.83 \cdot 10^{-1}$	1.03		$3.74 \cdot 10^{-2}$	1.08		$-2.34 \cdot 10^{-1}$	1.08	**
C2	$4.58 \cdot 10^{-3}$	1.00	*	$9.89 \cdot 10^{-4}$	1.03	***	$-1.02 \cdot 10^{-1}$	1.07		$-5.90 \cdot 10^{-1}$	1.07	***
C3	$8.08 \cdot 10^{-3}$	1.00	***	$1.44 \cdot 10^{-1}$	1.03	***	$-1.98 \cdot 10^{-1}$	1.07	**	$-5.99 \cdot 10^{-1}$	1.07	***
C4	$8.25 \cdot 10^{-3}$	1.00	***	$2.76 \cdot 10^{-1}$	1.03	***	$-2.21 \cdot 10^{-1}$	1.08	**	$-3.87 \cdot 10^{-1}$	1.09	***
C5	$4.09 \cdot 10^{-3}$	1.00	.	$2.09 \cdot 10^{-1}$	1.03	***	$-2.13 \cdot 10^{-1}$	1.09	**	$-9.27 \cdot 10^{-1}$	1.09	***

ages. Moving onto to cluster 2, these drivers have the second level of volatility. Drivers in this range tend to have claims anywhere between 21 to 7511,with standard deviation of 1047, and a mean of 1823. The ages of these drivers are between 19 to 62 with a mean age of 42. One can interpret this cluster as middle aged drivers. The claims of these drivers have higher volatility than the previous cluster, in this case age is very concentrated together with only a standard deviation of just 7 years, in contrast to the previous cluster's standard deviation of 15. Proceeding to the third cluster, its volatility in claims is greater. Drivers in this cluster have claims anywhere between 2 to 301300, with a mean of 3109, and a standard deviation of 13468. The drivers in this cluster are much older than the previous one with a mean age of 52 years and a standard deviation of 12.6. Finally cluster 4 denotes the level of highest volatility. Claims in this cluster of drivers reach the highest cost of 2037000, a mean of 5043, and a standard deviation of 59750.99. Their ages tend to

be much younger, with a mean age of 26, and a standard deviation of just 4. This is fairly standard for car insurance since younger drivers tend to take more risks.

Coefficients of clustered results are used to calculate premiums in car insurance. Table 5 shows the coefficients of the fitted model. The significance codes are defined as ≈ 0 (***), 0.001 (**), 0.01 (*), 0.05 (.) pertaining to the p value of the specific coefficient. In each cluster significance varies but overall the majority of coefficients are significant. An interpretation of the results is as follows. The first cluster (V1) has significant coefficients of Region (R#), Power (P#), Car age (C#), and Driver age. We see that the sign of coefficients change depending on the relative center of each cluster. The comparison between driver age in V1 and V4 is a perfect example. The coefficient of driver age is positive in V1 which shows that with larger age, claims tend to go up. Comparing that with last cluster, (V4). The driver age coefficient is negative. This is due to the fact that ages in V4 are very young (mean age of 25). Thus as age increases drivers tend to take less risks and become more responsible, which would decrease cost of claims. We see the same interpretation with Density. Clusters V2, V3, and V4 have density as a significant coefficient. However, the sign of the coefficient changes to negative in V4, while V2 and V3 have positive coefficients. Again this due to the relative centres of the cluster indicate both the magnitude and sign of the coefficient.

To summarize, the drivers have been clustered into four categories with distinct characteristics outlined in Table 4. We have seen how using the results from GL-TCWM, one can create an insurance model based on clustering algorithms with various levels of risk represented in each cluster. An interesting result was that cluster 1 contradicted the usual understanding that young people take more risks. The GL-TCWM method found a group that was the clear majority of drivers, in which the volatility of their claims was extremely low regardless of Driver Age, or Density. This finding shows that GL-TCWM may potentially find unique models that are otherwise hidden within the data.

4.2.3 Modeling Claims Frequency

In this section, we model frequency of the French motor claims. We consider the covariates log-density, driver age, car age, and a three-class grouping of the power of a car labelled Power F. The choice of covariates stems from the previously modelled single component ZIP (Charpentier, 2014). For the purposes of computational feasibility, only claims from the largest populated insurance region (R24) had been selected. The extension into multiple components is modelled with the linear formula

$$\text{Claim Nb} \sim \text{Driver Age} + \text{Log Density} + \text{Car Age} + \text{Power F}.$$

After fitting the model, the size of each cluster is noted. The GL-TCWM has chosen 3 components as the best model to represent the data. The size of each cluster is displayed in Table **??**. Attention is brought to cluster 2 which holds nearly 67% of the total population. Cluster 3 holds the fewest amount of drivers with merely 2.4% of the total population. Table 4 shows an in-depth analysis of Driver Age and Claim Number as shown in 3. Cluster 1 has the youngest drivers of the whole population with a mean age of 29.39 and a standard deviation of 4.87. Cluster 2 has a relatively older age group with a mean age of 36.90, and a standard deviation of 1.20. Finally Cluster 3 shows the oldest age group of drivers with a mean age of 54.25 and a standard deviation of 11.66. However when looking at the relative proportion of claims for each cluster as shown in Table 6. One notices that the middle age group has a higher proportion of non-zero claims. Relative to the other clusters, the middle age group has a non-zero claim proportion of 17.25 %. This is bolded in the Table 6 to show the difference of cluster 2 in comparison to 1, and 3.

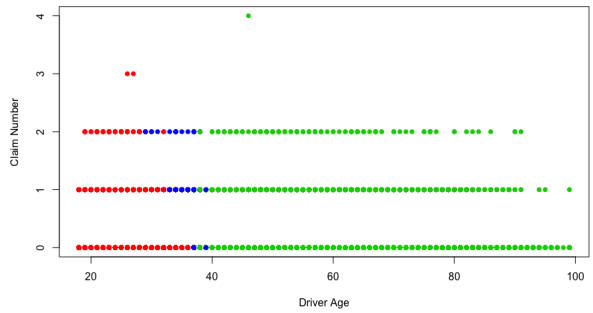

Figure 4: Claim number versus driver age, with partitions denoted by colour.

Table 6: Claim Nb analysis of clusters

Cluster	Claim Nb	Counts	Proportion (%)
1	0	48172	96.98 %
	1	1432	2.90 %
	2	64	0.12 %
	3	2	$\approx 0\%$
2	0	3262	**82.75 %**
	1	658	**16.69 %**
	2	22	**0.56 %**
3	0	102905	96.18%
	1	3963	3.71 %
	2	120	0.11 %
	4	1	$\approx 0\%$

Table 7: Size of clusters for modelling claims

1	2	3
49670	106989	3942
Red	Green	Blue

Table 8: Driver Age analysis of clusters.

Cluster	Minimum	Maximum	Mean	σ
1	18	38	29.39	4.87
2	28	39	36.90	1.20
3	38	99	54.25	11.66

The significance of the codes are defined as ≈ 0 (***), 0.001 (**), 0.01 (*), 0.05 (.) pertaining to the p value of the specific coefficient. For each cluster the significance of the coefficients vary greatly. In cluster 1, the count parameters (β_{xn}) for the Intercept, Driver Age and Log-Density are highly significant (***). The zero parameters β_{xz} are all highly significant with the exception of the power group GH (β_{5z}). After comparing the zero-inflated model for cluster 2 with a reduced Poisson model, the Vuong statistic had chosen the simpler Poisson model shown in Table 9. The Poisson model shows significance for the Intercept, Car Age, and Log-Density. Cluster 3 shows signficance for both zero and count models. The Intercept, Driver Age, and Log Density for both models show high significance. In summary the drivers have been clustered into three categories outlined in Table 9. From the coefficients, one can generate a premium plan tailored to the specific classifications of drivers using GL-TCWM.

Table 9: Table of coefficients for each cluster

Par	Est.	Std. Err	Z Val.	P	Est.	Std. Err	Z Val.	P	Est.	Std. Err	Z Val.	P
β_{0n}	-1.621	0.213	-7.629	***	-3.581	0.102	-35.109	***	6.973	0.923	7.553	***
β_{1n}	-0.075	0.006	-11.631	***	0.000	0.001	0.012		-0.220	0.028	-7.773	***
β_{2n}	-0.006	0.005	-1.181		-0.015	0.003	-5.481	***	0.003	0.009	0.325	
β_{3n}	0.113	0.015	7.589	***	0.085	0.010	8.884	***	0.103	0.030	3.401	***
β_{4n}	-0.065	0.086	-0.752		0.094	0.054	1.611		-0.087	0.153	-0.570	
β_{5n}	0.081	0.092	0.875		0.074	0.058	1.287		-0.006	0.167	-0.039	
β_{0z}	-228.080	55.429	-4.115	***					-106.767	6.873	-15.534	***
β_{1z}	7.601	1.832	4.149	***					3.059	0.194	15.782	***
β_{2z}	-0.313	0.128	-2.447	*					0.017	0.019	0.921	
β_{3z}	-4.309	1.002	-4.303	***					-1.133	0.087	-12.967	***
β_{4z}	7.891	2.105	3.750	***					-0.094	0.313	-0.301	
β_{5z}	-1.267	1.346	-0.941						0.096	0.334	0.288	

5 Simulation Study

Two simulation studies are conducted to determine the validity of transformation and the effectiveness of the Bernoulli-Poisson partitioning method. The first section outlines the need for transformation in the covariates.

The second section shows the classification accuracy and other relevant analysis for the Bernoulli-Poisson method.

5.1 Simulation Study - Transformation

In this section, we show how the proposed methodology works for different simulation settings. The simulation study was generated based on the regression coefficients of the **CASdataset** used in the previous section. The aim of the simulation study was to test the accuracy and ability of both GL-TCWM and CWM to return estimates of true parameters when one or more of the covariates is lognormal and the other two are Gaussian. This was designed to test both functions in the event when one of the covariates is non-Gaussian. The motivation behind this is fact is that many covariates used in insurance are likely to come from non-Gaussian distribution. Thus this was aimed to test the relevancy of CWM, which treats all covariates as Gaussian.

We define Model 1 as the base line model in which the coefficients were generated for **CASdataset** and reported in upper portion of Table 2. These coefficients were then rounded and treated as true parameters. A simulation with three GLM mixture components was then generated around these true parameters in which the third covariate X_3 was lognormal. Stemming from this, both CWM and GL-TCWM were run. The GL-TCWM treats X_3 as a lognormal covariate which then applies the relevant transformation.

The results for Model 1 were summarized in upper portion of Table 3 based on the performance of the GL-TCWM approach. The simulation was run 1000 times. We reported the percentage of runs for each predictor and the corresponding intercept in each mixture component under the assumption of 5% error. For example, predictor X_2 in the component 2 of Model 1 reported 90.10% accuracy. This means that 90.1% of the time the true parameter was estimated within 5% error. In this setting, predictor X_1 in the second component was insignificant in the real data set. The purpose of including this parameter in Model 1 was to test the sensitivity of GL-TCWM for insignificant predictors. In this case, the result of zero is underlined and it means that it has no influence on the response variable in this simulation.Further, we created Models 2, 3, 4 and 5 by altering the parameters of Model 1 by $+30\%$, -30%, $+50\%$, and -50% accordingly and keeping the second covariate of the second component as an insignificant predictor form the **CASdataset** model. This was done to test the accuracy of GL-TCWM to the sensitivity of coefficients. Based on the results in Table 3, we can see that GL-TCWM performs well for all simulation settings.

Table 4 provides the summary of the results when CWM was used in the analysis of the same models considered in Table 3. It is not surprising to see that barely any of the simulation runs estimated correctly all

Table 10: GL-TCWM vs CWM Accuracy: Covariate X_3 is treated as log-normal, the rest are Gaussian covariates. The transformation of X_3 is considered.

Model	Component	Intercept	X_1	X_2	X_3	Intercept	X_1	X_2	X_3
1	1	93.00%	90.10%	93.00%	93.10%	0.00%	0.00%	0.00%	0.00%
	2	90.10%	0.00%	90.10%	90.10%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.10%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
2	1	89.80%	89.20%	89.80%	89.80%	0.00%	0.00%	4.60%	0.00%
	2	89.20%	0.00%	89.20%	89.20%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.20%	1.70%	0.00%
3	1	100.00%	100.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	2	100.00%	0.00%	100.00%	100.00%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
4	1	88.60%	86.80%	88.60%	87.00%	0.00%	0.00%	0.00%	0.00%
	2	86.90%	0.00%	86.90%	86.90%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.00%	0.00%	0.00%
5	1	85.90%	84.90%	85.60%	85.90%	0.00%	0.00%	0.00%	0.00%
	2	85.00%	0.00%	84.90%	84.90%	0.00%	0.00%	0.00%	0.00%
	3	99.20%	99.20%	99.20%	99.20%	0.00%	0.20%	10.90%	0.00%

parameters as most of the results are zero. This means that the performance of CWM approach is poor in presence of one non-Gaussian covariate which in this case is a log-normal covariate. Similarly to Table 3, Table-4 shows the underlined results pointing to insignificant predictors.

Table 11: GL-TCWM results: the summary of MSE for all parameters used in five models. The covariate X_3 is treated as log-normal a nd the rest are Gaussian. These results correspond to those in Table 3.

Model	Component	β_o	MSE(β_o)	β_1	MSE(β_1)	β_2	MSE(β_2)	β_3	MSE(β_3)
1	1	1028	(11.353)	0.03	(0.00)	3.5	(0.00)	-380	(0.09)
	2	1600	(0.000)	-0.01	(0.00)	1.5	(0.00)	-250	(0.00)
	3	40000	(0.035)	-6.00	(0.00)	-305	(0.00)	1100	(0.47)
2	1	1350	(0.167)	0.04	(0.00)	4.5	(0.00)	-500	(0.03)
	2	2080	(0.001)	0.04	(0.00)	2.0	(0.00)	-325	(0.00)
	3	52000	(0.012)	-8.00	(0.00)	450	(0.00)	14300	(0.01)
3	1	720	(0.001)	0.02	(0.00)	2.5	(0.00)	-266	(0.00)
	2	1100	(0.008)	0.00	(0.00)	1.1	(0.00)	-17511	(0.00)
	3	28000	(0.002)	-4.20	(0.00)	245	(0.00)	7700.	(0.00)
4	1	1650	(13.056)	0.05	(0.00)	5.3	(0.00)	-570	(0.00)
	2	2400	(0.000)	-0.01	(0.00)	2.3	(0.00)	-375	(0.00)
	3	60000	(0.051)	-9.00	(0.00)	-457	(0.00)	16500	(0.00)
5	1	500	(1.115)	0.02	(0.00)	2.0	(0.00)	-190	(0.05)
	2	800	(0.003)	0.00	(0.00)	0.8	(0.00)	-120	(0.00)
	3	20000	(0.000)	-3.00	(0.00)	-150	(0.00)	5500	(0.00)

Table 5 provides the summary of Mean Squared Errors (MSE) of each parameter of the models in Table 3 estimated via 1000 simulation runs. The MSE is computed using the following formula $\text{MSE}(\beta_i) = \frac{\sum_i^n (\beta_i - \hat{\beta}_i)^2}{n}$. The MSEs related to the predictor variables for all models and their corresponding components are about zero indicating that GL-TCWM approach performs well. This is also a result of having a small size coefficients.

Table 12: CWM results: the summary of MSE for all parameters used in five models. All three covariates are treated as Gaussian. These results correspond to those in Table 4.

Model	Component	β_o	MSE(β_o)	β_1	MSE(β_1)	β_2	MSE(β_2)	β_3	MSE(β_3)
1	1	1028	(·)	0.03	(·)	3.5	(·)	-380	(·)
	2	1600	(·)	-0.01	(·)	1.5	(·)	-250	(·)
	3	40000	(·)	-6.00	(·)	-305	(·)	1100	(·)
2	1	1350	(·)	0.04	(·)	4.5	(·)	-500	(·)
	2	2080	(·)	0.04	(·)	2.0	(·)	-325	(·)
	3	52000	(·)	-8.00	(0.006)	450	(44.1)	14300	(·)
3	1	720	(·)	0.02	(·)	2.5	(·)	-266	(·)
	2	1100	(65.814)	0.00	(·)	1.1	(·)	-17511	(·)
	3	28000	(·)	-4.20	(·)	245	(·)	7700.	(·)
4	1	1650	(·)	0.05	(·)	5.3	(·)	-570	(·)
	2	2400	(·)	-0.01	(·)	2.3	(·)	-375	(·)
	3	60000	(·)	-9.00	(·)	-457	(·)	16500	(·)
5	1	500	(·)	0.02	(·)	2.0	(·)	-190	(·)
	2	800	(·)	0.00	(·)	0.8	(·)	-120	(·)
	3	20000	(·)	-3.00	(0.003)	-150	(4.7)	5500	(·)

Table 6 provides the summary of Mean Squared Errors (MSE) of each parameter of the models in Table 4 estimated via 1000 simulation runs. In contrary to the results reported in Table 5, these results in Table 6 are significantly different. We can observe that the MSEs for most of the Models and their corresponding components are not generated at all and as such they are shown as $(\cdot)$. This is not surprising because Table 4 shows the accuracy of CWM is not good when attempting to model non-Gaussian predictors as Gaussian. In summary, our simulation results showed good performance of GL-TCWM approach in modeling non-Gaussian covariates. More specifically, these results show high accuracy when covariates are log-normal. In contrary, CWM fails to estimate parameters accurately when the Gaussian assumption is violated.

5.2 Simulation Study - Bernoulli-Poisson Partitioning

In this section we show how the Bernoulli-Poisson partitioning (BP) method behaves under different conditions. The components were genereated under similar coefficients taken from the **CASDatasets**

package. The coefficients were rounded and treated as true parameters to which data was generated from. The mean and standard deviation of the covariates within each component was also taken into account when generating data. The first simulation examines the performance of the GL-TCWM model for classification. We generate three components each with sample size $N = 1000$ for a total of 3000 simulated points. The model generated is similar to the mean and standard deviations of Table 8. Consider three simulated covariates and the following GLM model

$$\text{SimClaims} \sim \text{SimDriverAge + SimLogDensity + SimCarAge.}$$

Here the GL-TCWM model is fitted to the simulated data and used to classify into three components. The misclassification rate is calculated by the proportion of true labels placed in other components by the GL-TCWM model. The results of the simulation is based on the generated dataset are presented in Table 8. The total misclassification rate is 1.8% and the majority of misclassified components are between components two and three.

Table 13: Misclassfication rate and label comparison of generated data.

True Labels	Classified			Misclassification Rate (%)
	1	2	3	
1	992	3	5	0.80
2	0	990	10	1.00
3	15	20	965	3.50
Overall Misclassification Rate				1.80 %
Average Purity				98.23 %
Adjusted Rand Index				0.9479

$$n_{ij} = \text{ across diagonal }, \qquad a_i = \text{ row sums }, \qquad b_j = \text{ column sums}$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \qquad AP = \frac{1}{N}\sum_i n_{ij}$$

The experiment is expanded further to show how Bernoulli-Poisson partitioning behaves over 1000 runs and under two different conditions. The first condition is defined as follows. The mean and standard deviations are taken as given by the estimated ZIP components from the **CASDataset**. The second condition involves adjusting the means of two of the covariates so they are closer to each other. The goal is to show that the BP-method holds its use even when means among covariates are close. Conditions are divided into two

categories. N is considered normal, where the covariate means are taken directly from the sample data. C is considered to be "close", where the covariate means are manipulated so that they are closer to each other within some degree. This is a common problem in classification where if the means among two different components are close, then misclassification rate increases [Lim et al. (2014)]. Experiment 2 defines the use of 3 different partitioning methods to initialize a zero-inflated model. Poisson method assumes that the presence of non-zeros will provide a better partitioning of the data-set. Bernoulli assumes that the presence of excess zeros will determine the best partitioning of the data-set . Finally the BP-Method assumes that both methods are weighed equally and therefore both must be taken into account when partitioning the dataset. The mean and standard deviation of each measurement is provided in Table 14.

Table 14: Experiment 2: mean and standard deviations for each statistic comparing each method.

Type	Condition	Poisson (σ)	Bernoulli (σ)	BP-Method (σ)
Misclassification Rate	N	1.70% (6.00)	1.60% (6.00)	1.10% (0.02)
	C	5.00% (7.00)	6.00% (2.00)	7.00% (4.00)
Average Purity	N	98.87% (2.00)	98.91% (2.25)	99.18% (0.81)
	C	95.38% (4.00)	94.55% (1.00)	96.95% (0.48)
Adjusted Rand Index	N	0.9662 (0.07)	0.9677 (0.07)	0.9729 (0.0217)
	C	0.8706 (0.08)	0.8366 (0.04)	0.8538 (0.0453)

Several findings are concluded from Table 14. Under condition N, the BP method shows better performance in error and is found to be less sensitive than other methods with an error rate of 1.10% and a standard deviation of 0.02%. Further findings show that when condition C is imposed then Bernoulli has better performance in terms of findings. The ARI shows good measurements overall however the BP-Method under condition N has a very good ARI with a small standard deviation. The Average Purity of the BP-Method is the best out of all other methods, which is relevant to estimating coefficients accurately for the ZIP optimization.

6 Conclusion

In this paper, we extend the class of generalized linear mixture CWM models by accomplishing two main goals. First, we propose the methodology that allows for continuous covariates to follow a non-Gaussian distribution. Imposing Gaussian distribution on a skewed data may result in an suboptimal model fit. Second, we propose a new Poisson CWM methodology that uses Bernoulli-Poisson partitioning and allows for implementation of zero-inflated Poisson CWM model. We call our proposed model class GL-TCWM which reflects two extensions made to the existing CWM class of models.

The GL-TCWM models allows for great applications in predictive modeling of insurance claims by overcoming a few limitations of the current CWM models. Zero-inflated GL-TCWM allows for finding clusters within claims frequency which is an important information in risk classification and modeling of claims frequency. Further, some insurance rating variables used in the predictive modeling of severity claims may not strictly follow Gaussian assumptions, e.g. driver's age or car age (treated as continuous covariates). An adequate transformation can be considered on the continues covariates to relax current assumptions and improve the model fit. We demonstrated that if there is a need for transformation, a Lognormal transformation can be considered easily to improve the model fit.

The results of our extensive simulation study showed the excellent performance of the proposed model in case of modeling non-Gaussian covariates. We found that current CWM model fails to estimate the parameters accurately when the Gaussian assumption is violated. The GL-TCWM shows significant improvement in the model fit over the CWM model based on AIC and BIC criteria. We also tested Bernoulli-Poisson partitioning of zero-inflated GL-TCWM under different conditions and found that our proposed partitioning method has a very low misclassification rate, high average purity, and high average rand index.

Our approach is relevant to the actuarial pricing and risk management when current practices are based on implementation of various GLM models. Further extension of this work may incorporate Bayesian setting by exploring different assumption on informative and noninformative priors (see Ibrahim and Laud (1991)). By utilizing these techniques actuaries will be able to make inferences from posterior distributions of model parameters and from predictive distributions in order to improve pricing and risk management of the insured portfolio.

7 Appendix

7.1 Proof of Transformation

Proof. Consider the univariate Log-normal distribution with $x \in \mathbb{R}^+$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, then the density of x is defined as

$$\mathcal{LN}(x;\mu,\sigma)dx = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$$

We then consider that the change of variable property must conserve differential probability. In doing so we can write the density of x as a Gaussian form.

$$\mathcal{LN}(x;\mu,\sigma)dx = \frac{\mathcal{N}(\ln x;\mu,\sigma)}{x}dx = \mathcal{N}(\ln x;\mu,\sigma)\frac{d\ln x}{dx}dx = \mathcal{N}(\ln x;\mu,\sigma)d\ln x$$

Thus Log-normal distribution can be written as a Gaussian form.

$$\mathcal{N}(\ln x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] = \mathcal{LN}(x; \mu, \sigma) * (x) \tag{7.1}$$

Using the change of variable property we can extend to a Log-normal multivariate case where $\boldsymbol{x} \in \mathbb{R}^{+p}, \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} \in \mathbb{R}^{+p}$, where $p \in \mathbb{N}$ denoting the p-dimensional space, then the density of $\boldsymbol{x}$ is written as:

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\prod_{i=1}^{N} x_i)|\boldsymbol{\Sigma}|(2\pi)^{\frac{p}{2}}} \exp\left[-\frac{1}{2}(\ln(\boldsymbol{x}) - \boldsymbol{\mu_g})\boldsymbol{\Sigma_g^{-1}}(\ln(\boldsymbol{x}) - \boldsymbol{\mu})\right]$$

Similarly by (7.1), $f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written in a multivariate Gaussian form as follows

$$f(\ln \boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{i=1}^{N} x_i. \tag{7.2}$$

Hence, the transformation function is defined as

$$\phi\left(f(\boldsymbol{x}; \boldsymbol{\theta})\right) := f(\boldsymbol{x}; \boldsymbol{\theta}) \prod_{i=1}^{N} x_i \tag{7.3}$$

where $f(\boldsymbol{x}, \boldsymbol{\theta})$ is a density function with parametrized vector $\boldsymbol{\theta}$. This concludes the existence of a bijective map from a Log-normal to a Gaussian form.

□

References

Bermúdez, L., Karlis, D., 2012. A finite mixture of bivariate poisson regression models with an application to insurance ratemaking. Computational Statistics and Data Analysis 56 (12), 3988–3999.

Brown, G., Buckley, W., 2015. Experience rating with poisson mixtures. Annals of Actuarial Science 9 (02), 304–321.

Charpentier, A., 2014. Computational Actuarial Science with R. CRC press.

Czado, C., Kastenmeier, R., Brechmann, E., Min, A., 2012. A mixed copula model for insurance claims and claim sizes. Scandinavian Actuarial Journal 4, 278–305.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society B 39, 1–38.

Ezzahid, E., 12 2012. Poisson regression and zero-inflated poisson regression: application to private health insurance data 2.

Frees, E.W., D. R., Meyers, G. e., 2014. Predictive modeling applications in actuarial science. Cambridge University Press 1.

Frees, E., Lee, G., Yang, L., 2016. Multivariate frequency-severity regression models in insurance. Risks 4 (1), 4.

Frees, E., Wang, P., 2006. Copula credibility for aggregate loss models. Insurance: Mathematics and Economics 38 (2), 360–373.

Garrido, J., Genest, C., Schulz, J., 2016. Generalized linear models for dependent frequency and severity of insurance claims. Insurance: Mathematics and Economics 70, 205–215.

Gershenfeld, N., 1997. Nonlinear inference and cluster-weighted modeling. Annals of the New York Academy of Sciences 808 (1), 18–24.

Gershenfeld, N., 1999. The nature of mathematical modeling. Cambridge university press.

Gershenfeld, N., S. B., Metois, E., ., 1999. Cluster-weighted modelling for time-series analysis. Nature 397 (67171), 329–332.

Ibrahim, J., Laud, P., 1991. On bayesian analysis of generalized linear models using jeffreys's prior. Journal of the American Statistical Association 86 (416), 981–986.

Ingrassia, S., M. S., Punzo, A., 2014. Model-based clustering via linear cluster-weighted models. Computational Statistics & Data Analysis 71, 159–182.

Ingrassia, S., M. S., Vittadini, G., 2014. Local statistical modeling via a cluster-weighted approach with elliptical distributions. Journal of classification 29 (3), 363–401.

Ingrassia, S., P. A. V. G., Minotti, S., 2015. The generalized linear mixed cluster-weighted model. Journal of Classification 32 (1), 85–113.

Krämer, N., Brechmann, E., Silvestrini, D., Czado, C., 2013. Total loss estimation using copula-based regression models. Insurance: Mathematics and Economics 53 (3), 829–839.

Lambert, D., 02 1992. Zero-inflated poisson regression, with an application to defects in manufacturing 34, 1–14.

Lee, S. C. K., Lin, X. S., 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. North American Actuarial Journal 14 (1), 107–130.

Lim, H., Li, W., Yu, P., 03 2014. Zero-inflated poisson regression mixture model 71, 151?158.

McLachlan, S., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, Hobuken, NJ.

McNicholas, P., Murphy, T., McDaid, A., Frost, D., 2010. Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. Computational Statistics Data Analysis 54 (3), 711 – 723, second Special Issue on Statistical Algorithms and Software.
URL http://www.sciencedirect.com/science/article/pii/S0167947309000632

Miljkovic, T., Grün, B., 2016. Modeling loss data using mixtures of distributions. Insurance: Mathematics and Economics.

Punzo, A., Ingrassia, S., ., 2014. Parsimonious generalized linear Gaussian cluster-weighted models. Springer International Publishing.

Shi, P., Feng, X., Ivantsova, A., 2015. Dependent frequencyseverity modeling of insurance claims. Insurance: Mathematics and Economics 64, 417–428.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., Lin, S., 2015. Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. ASTIN Bulletin 45 (3), 729–758.

Wedel, M., 2002. Concominat variables in finite mixture modeling. Statistica Neerlandica 56 (3), 362–375.

Wedel, M., De Sabro, W., 1995. A mixture likelihood approach for generalized linear models. Journal of Classification 12 (3), 21–55.